Stacking & LSM Namespacing Redux

Linux Plumbers Container MC 2018

Casey Schaufler – Intel

John Johansen - Canonical

Linux Security Modules (LSM)

- Provide security
- Often MAC but not necessarily
- Kernel provides security
  - Hooks
    - Located at security decision points
    - All security relevant info available
    - Race free
  - Security field in various objects

- selinux, smack, apparmor, tomoyo, IMA/EVM, loadpin, yama
- proposed: LSMs: LandLock, CaitSith, Checmate, HardChroot, PTAGS,
  SimpleFlow, SafeName, WhiteEgret, shebang, S.A.R.A.

Use Cases

- LSM enabled in container but not on Host
  - ChromeOS running Android SELinux container
  - Virtual smart phone env (Cells/Cellrox), multiple android instances
  - Thin linux host (clear linux)
- system container
  - lxd. run Ubuntu (apparmor) container on rhel (selinux) host
- application confinement
  - snap using apparmor running on fedora (selinux base system)
  - Docker
  - flatpak

Problem

The LSM is not Namespaced

LSM Namespacing

- Just Create an LSM Namespace!
- Presented & Discussed idea at Linux Plumbers 2017
  - Not enough semantic info at LSM layer
  - Some LSMs don’t want to be “namespaced”
    - Want to bound container
    - No generic Solution
  - Real work needs to be done in security modules

Namespacing the LSMs

Requirements

- Not every LSM has the same requirements
- System level confinement (confine the container)
  - eg selinux using MCS label per container
  - do NOT want either OR mediation
    - ie. selinux mediating tasks outside
    - container using different LSM not confined by selinux
- Application level confinement
  - Not every LSM supports
- Dependent Components Need support (audit, ...)

Audit

- Want ContainerID
  - But …
- Dependency of LSMs (apparmor, selinux, smack, ima)
- Not Namespaced
- Single Set of Rules
- Single daemon registration

Audit LSS16: Conclusion

- Auditd ok with MNT, UTS, IPC, CGRP ns
- NET ns ok for now
  - Will need audit_pid/portid per USER ns
- PID ns ok for now for audit user messages
  - Will need translation per PID ns
- Auditd per USER ns wanted for containers
- NamespaceID vs. Audit ContainerID
- Need audit log aggregation by container orch

AuditID

- U64
- containers can't be universally identified by namespace (sub)set
- audit daemon won't be tied to any namespace
- netNS needs list of possible IDs responsible for net events
- child inherits parent's ID
- allow multiple audit daemons
  - each will have its own queue and ruleset
  - auxiliaries can't influence host

SELinux NS

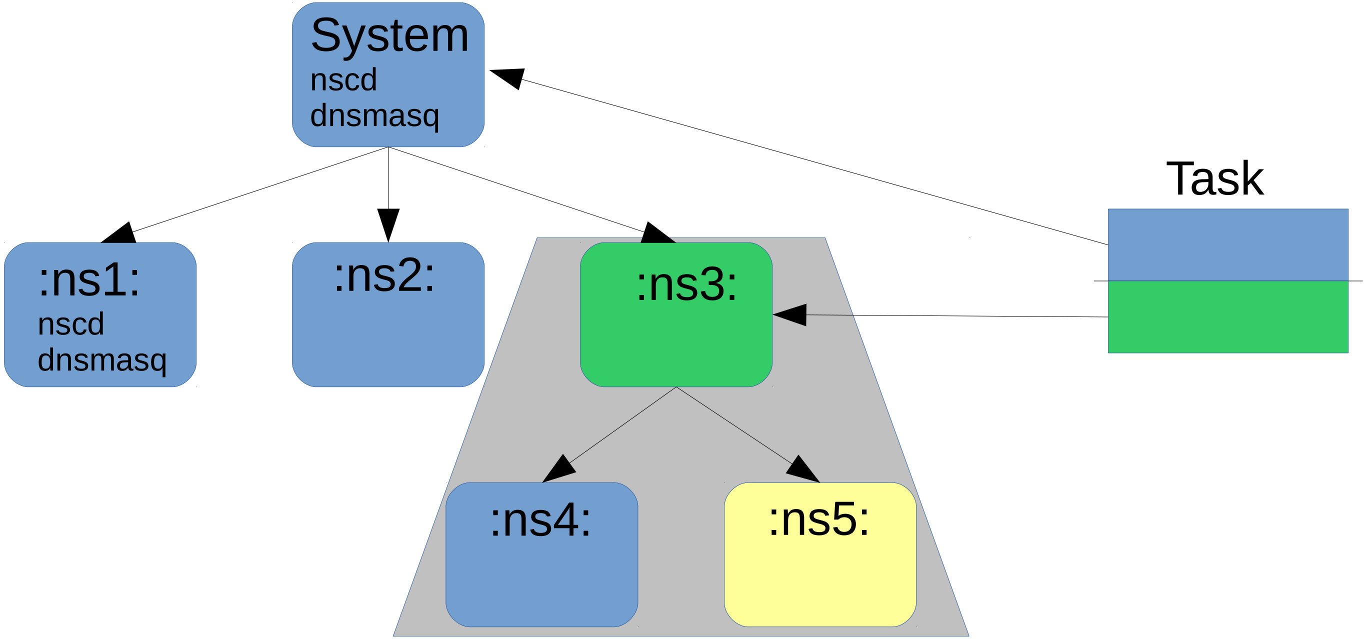
- Adds per-namespace selinuxfs instances
  - unshare mount ns and mount new selinuxfs
- Move AVC into namespace
- Add per-namespace support for kernel objects
- Write to selinuxfs unshare node to instantiate
- On Disk Inodes store all each NS label
- NS
  - Track nesting
  - Bounded enforcement

SELinux prototype

```
echo 1 > /sys/fs/selinux/unshare
unshare -m -n
umount /sys/fs/selinux
mount -t selinuxfs none /sys/fs/selinux
load_policy
runcon unconfined_u:unconfined_r:unconfined_t:s0:c0.c1023 /bin/bash
setenforce 1
```

AppArmor

- Namespaced
- Stacked
- Virtualized fs



AppArmor Problems

- Namespacing
  - mount, network, user, .. pita
    - Need more infrastructure
- Securityfs
  - can’t mount multiple instances need to bind mount
- Still only AppArmor in AppArmor containers

IMA

- Really wants ContainerID
- Prototype
  - IMA Audit
  - Virtualized IMA fs interface
- EVM
  - Problems with ns xattr storage

Other LSMs

- Smack
  - Prototype namespace from a few years ago
- Yama
- Loadpin

- Landlock
- Sara

Stacking the LSMs

Stacking Enablement

- LSMs enabled at boot
  - Reserve space for kernel objs
  - Infrastructure manages life time
  - Register hooks
- New kernel param
  - LSM=

Making Stacking Work

Problem

- Userspace Interfaces
  - /proc/pid/attr/*
  - SO_PEERSEC

Fix

- Virtualize – per task default LSM

Making Stacking Work

Problem

- Userspace Interfaces
  - /proc/pid/attr/*
  - SO_PEERSEC

Fix

- Virtualize – per task default LSM
- Interface to set default LSM

Making Stacking Work

Problem

- Userspace Interfaces
  - /proc/pid/attr/*
  - SO_PEERSEC

Fix

- Virtualize – per task default LSM
- Interface to set default LSM
- New versions of interfaces
  - /proc/pid/attr/apparmor/*
  - /*proc*/pid/attr/smack/*
- …

Making Stacking Work

Problem

- Userspace Interfaces
  - /proc/pid/attr/*
  - SO_PEERSEC
- Networking
  - secids

Fix

- Virtualize – per task default LSM
- Interface to set default LSM
- New versions of interfaces
  - /proc/pid/attr/apparmor/*
  - /*proc*/pid/attr/smack/*
- …
  - Dynamically compose & remap

Making Stacking Work

Problem

- Userspace Interfaces
  - /proc/pid/attr/*
  - SO_PEERSEC

- Networking
  - secids
  - secmark

Fix

- Virtualize – per task default LSM
- Interface to set default LSM
- New versions of interfaces
  - /proc/pid/attr/apparmor/*
  - /*proc*/pid/attr/smack/*
- …

  - Dynamically compose & remap
  - Extend to support multiple LSMs

Making Stacking Work

Problem

- Userspace Interfaces
  - /proc/pid/attr/*
  - SO_PEERSEC

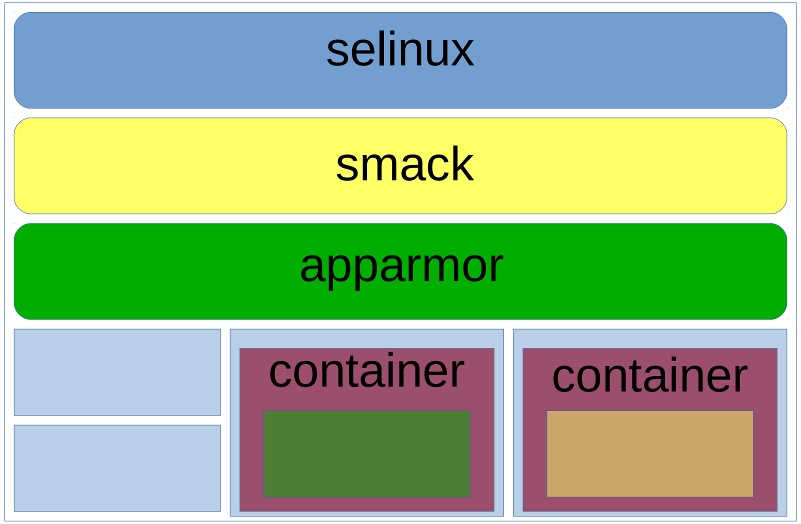
- Networking
  - secids
  - Secmark
  - Netlabel cipso/calypso/xfrm

Fix

- Virtualize – per task default LSM
- Interface to set default LSM
- New versions of interfaces
  - /proc/pid/attr/apparmor/*
  - /*proc*/pid/attr/smack/*
- …

  - Dynamically compose & remap
  - Extend to support multiple LSMs
  - Only 1 LSM may claim and use

Current Situation with Stacking



References & Thanks

Linux Audit – Moving Beyond Kernel Namespaces to Audit Container IDs
    Richard Guy Briggs, Linux Security Summit NA 2018

Namespacing in SELinux
    James Morris, Linux.conf.au 2018

Security Module Stacks that Don’t Fall Over
    Casey Schaufler, Linux Security Summit EU 2018

Overview and Recent Developments: Linux Integrity
    Mimi Zohar, Linux Security Summit EU 2018

Overview and Recent Developments: AppArmor
    John Johansen, Linux Security Summit EU 2018